Санкт-Петербургский Политехнический Университет им. Петра Великого

Высшая школа прикладной математики и вычислительной физики

Отчёт по лабораторной работе №1 по дисциплине “Теория принятия экономических решений”

**Классификация многомерных объектов при наличии обучающей выборки**

Выполнил студент:

Мишутин Д. В.

Группа:

3630102/70301

Проверил:

К.ф.-м.н., доцент

Павлова Людмила Владимировна

Санкт-Петербург

2020 г.

Оглавление

[1 Постановка задачи 3](#_Toc56127654)

[2 Численные эксперименты 3](#_Toc56127655)

[2.1 Построение и тестирование классификатора с использованием модельных данных 3](#_Toc56127656)

[2.2 Построение и тестирование классификатора с использованием данных из репозитория 7](#_Toc56127657)

[3 Реализация 8](#_Toc56127658)

[4 Литература 8](#_Toc56127659)

[5 Приложения 9](#_Toc56127660)

# 1 Постановка задачи

Сгенерировать обучающие (ОВ) и тестовые (ТВ, представляющие смеси выборок, отличных от ОВ1 и ОВ2, с теми же параметрами) выборки размеров 100-1000 для модельных данных. Классифицировать объекты из каждой выборки на основе дискриминантного анализа (в нашем случае через бинарную классификацию), оценив параметры распределений получившихся выборок, константу , расстояние Махаланобиса и дискриминантную функцию (ДФ). Численные эксперименты провести на “хорошо” и “плохо” разделённых данных. Провести аналогичные исследования для данных из репозитория.

Пусть первый класс представлен обучающей выборкой 1 (ОВ1) объёма из нормального распределения .

Пусть второй класс представлен обучающей выборкой 2 (ОВ2) объёма из нормального распределения .

(векторы средних), (матрицы ковариаций). В нашем случае .

# **2 Численные эксперименты**

## 2.1 Построение и тестирование классификатора с использованием модельных данных

Пусть для “хорошо” разделённых данных параметры распределения будут следующими:

Для “плохо” разделённых:

1. После генерации были получены следующие выборочные характеристики для **ОВ**:

* Выборочные средние:

1. Для “хорошо” разделённых данных:
2. Для “плохо” разделённых:

Расчётная формула:

* Выборочные матрицы ковариации:

1. Для “хорошо” разделённых данных:
2. Для “плохо” разделённых:

Расчётные формулы:

* Оценка ДФ:

1. Для “хорошо” разделённых данных:
2. Для “плохо” разделённых:

Расчётная формула:

* Выборочные средние значения ДФ:

1. Для “хорошо” разделённых данных:
2. Для “плохо” разделённых:

Расчётная формула:

* Выборочная несмещённая дисперсия:

1. Для “хорошо” разделённых данных:
2. Для “плохо” разделённых:

Расчётная формула:

* Смещённая оценка расстояния Махаланобиса:

1. Для “хорошо” разделённых данных:
2. Для “плохо” разделённых:

Расчётная формула:

* Несмещённая оценка расстояния Махаланобиса:

1. Для “хорошо” разделённых данных:
2. Для “плохо” разделённых:

Расчётная формула:

2. Матрица соответствий для **ОВ**:

Расчётные формулы:

* Для “хорошо” разделённых данных:

|  |  |  |
| --- | --- | --- |
| Pred→  True ↓ | 1 | 2 |
| 1 | 700 | 0 |
| 2 | 0 | 300 |

Табличные оценки вероятности ошибочной классификации:

Теоретические оценки вероятности ошибочной классификации:

* Для “плохо” разделённых:

|  |  |  |
| --- | --- | --- |
| Pred→  True ↓ | 1 | 2 |
| 1 | 763 | 37 |
| 2 | 73 | 127 |

Табличные оценки вероятности ошибочной классификации:

Теоретические оценки вероятности ошибочной классификации:

3. Матрица соответствий для **ТВ**:

* Для “хорошо” разделённых данных:

|  |  |  |
| --- | --- | --- |
| Pred→  True ↓ | 1 | 2 |
| 1 | 150 | 0 |
| 2 | 0 | 150 |

Табличные оценки вероятности ошибочной классификации:

* Для “плохо” разделённых:

|  |  |  |
| --- | --- | --- |
| Pred→  True ↓ | 1 | 2 |
| 1 | 338 | 12 |
| 2 | 127 | 223 |

Табличные оценки вероятности ошибочной классификации:

4. Анализ полученных результатов:

Чем дальше друг от друга отстоят выборки (чем меньше расстояния между их средними и меньше дисперсии признаков), тем лучше бинарный классификатор справляется с их разделением, т. е. меньше вероятность ошибочной классификации.

С ростом мощностей выборок эмпирические оценки ошибочной классификации стремятся к теоретическим (вычисленным через функцию Лапласа).

## 2.2 Построение и тестирование классификатора с использованием данных из репозитория

Рассмотрим данные из репозитория (файл *german.data-numeric*). Данные представляют 1000 объектов (1000 резюме различных людей), относящихся к одному из двух классов. Каждый объект характеризует набор из 24 численных признаков, являющихся характеристиками материального, финансового, семейного положения и его трудоустройства.

К первому классу относятся объектов, ко второму - . Проведём разделение каждого класса на ОВ и ТВ в соотношении 75%:25% - 527:223 для ОВ и 173:77 для ТВ.

1. После разделения были получены следующие выборочные характеристики для **ОВ**:

* Выборочная несмещённая дисперсия:

Расчётная формула:

* Смещённая оценка расстояния Махаланобиса:

Расчётная формула:

* Несмещённая оценка расстояния Махаланобиса:

Расчётная формула:

1. Матрица соответствий для **ОВ**:

|  |  |  |
| --- | --- | --- |
| Pred→  True ↓ | 1 | 2 |
| 1 | 473 | 54 |
| 2 | 113 | 110 |

Табличные оценки вероятности ошибочной классификации:

Теоретические оценки вероятности ошибочной классификации:

1. Матрица соответствий для **ТВ**:

|  |  |  |
| --- | --- | --- |
| Pred→  True ↓ | 1 | 2 |
| 1 | 157 | 16 |
| 2 | 35 | 42 |

Табличные оценки вероятности ошибочной классификации:

1. Анализ полученных результатов:

Из оценок видно, что две популяции объектов выборки имеют значительную зону пересечения, из-за чего мы не можем качественно разделить два данных класса (доля ошибок на объектах второго класса близка к 50%).

Возможно лучше применить другой метод классификации (нелинейный) или повысить репрезентативность выборки.

# 3 Реализация

Была использована среда *IPython Notebook* (язык *Python 3.8.2*): модули *numpy* для генерации выборок с различными распределениями и математических расчётов, *pandas* - для хранения данных в таблицах, *math* – для математических вычислений; функции *norm.cdf* из модуля *scipy.stats* для вычисления функции Лапласа , *display* из модуля *IPython.display* для отображения pandas’ских таблиц в браузере и *train\_test\_split* из *sklearn.model\_selection* для разделения данных из репозитория на обучающие и тестовые.

# 4 Литература

[Основы работы с *numpy* (отдельная глава курса)](https://stepik.org/course/401)

Лекции по ТПЭР, “Дискриминантный анализ”, Павлова Л. В., 2020 г.

# 5 Приложения

[Код лабораторной](https://github.com/MeShootIn/economics-decision-theory/blob/main/lab_1/main.ipynb)